

АКАДЕМИЯ НАУК СССР, СИБИРСКОЕ ОТДЕЛЕНИЕ
ИНСТИТУТ МАТЕМАТИКИ

Анализ разнотипных данных
(Вычислительные системы, 99)
Сборник научных трудов
Новосибирск, 1983, с.44-50.

АЛГОРИТМ ЕСТЕСТВЕННОЙ КЛАССИФИКАЦИИ

ВИТЯЕВ Е.Е.

1. В настоящее время известно много принципов и алгоритмов построения классификаций [1-5]: на основе гипотезы компактности и различных мер близости в некотором пространстве; по эталонам - сходством с эталонами, преобразованиями эталонов, выбором эталонов или типичных представителей; по суперцелям (например, для последующего распознавания); по различным критериям качества классификации и функционалам качества; разделением смесей распределений и другие.

Тем не менее, не существует алгоритмов, основанных на принципах "естественной" классификации. Для пояснения смысла термина "естественная" классификация приведем критерии "естественности" классификации, которые давали естествоиспытатели.

1. "Естественной" является та, и только та классификация, которая выражает закон природы [6].

2. Критерий Уэвелла [6]: "Чем больше общих утверждений об объектах дает возможность сделать классификация, тем она естественней".

3. Критерий Любищева [6]: "наиболее совершенной системой (классификацией - Е.Е.) является такая, где все признаки объекта определяются положением его в системе. Чем ближе система стоит к этому идеалу, тем она менее искусственна, и естественной следует называть такую, где количество свойств объекта, поставленных в функциональную связь с его положением в системе, является максимальным (в идеале - это все его свойства)".

4. В работе [7], на основании анализа высказываний естествоиспытателей, выдвигается следующий критерий: "В многообразии объектов, образующих "естественную" классификацию, можно обнаружить два типа закономерностей:

1. соотношения, связывающие "короткое" описание архетипа, достаточное для диагностирования принадлежности объекта данному классу, с "полным" описанием. В сущности, это законы, позволяющие на основании принадлежности объекта некоторому естественному классу прогнозировать все его необходимые свойства;

2. правила показывающие как деформируются свойства объектов при переходе к смежным классам. Именно они гарантируют возможность переноса знаний с одного объекта на все принадлежащие данному классу и, несколько сложнее, на объекты смежных классов"

Мы сформулируем следующий принцип построения "естественных" классификаций [8]: "Разбиение объектов на классы должно производиться в соответствии с теми закономерностями, которым удовлетворяют объекты. Более точно этот принцип можно сформулировать следующим образом: Разбиение на классы должно производиться так, чтобы **объекты одного класса подчинялись одним и тем же закономерностям, объекты разных классов подчинялись разным группам закономерностей**. Объекты одного класса, кроме того,

должны обладать некоторой целостностью. **Целостность определим как взаимную согласованность закономерностей каждой группы по предсказанию** различных свойств объектов. У групп закономерностей могут быть общие закономерности, устанавливающие взаимосвязь признаков объектов из разных классов".

Данный принцип позволяет определить не только "естественную" классификацию, но и другие, связанные с ней понятия классиологии и соответствующие им методы анализа данных:

1. Определение групп закономерностей, описывающих классы;
2. Автоматическое формирование эталонов (идеальных представителей классов), как тех значений признаков, на которых закономерности максимально согласованы между собой по предсказанию значений признаков объектов класса, т.е. в максимальной степени проявляют свою целостность;
3. Определение иерархии классов;
4. Предсказание, заполнение пробелов в данных, нахождение ошибок в данных на основании предсказаний неизвестных (и известных) значений признаков по закономерностям, описывающим классы;
5. Распознавание, как предсказание принадлежности к образу;
6. Поиск уникальных объектов - объектов, составляющих самостоятельные классы;
7. Автоматическое формирование вторичных признаков, как выделение наиболее общих подклассов.

Разные определения понятий закономерности и взаимной согласованности должны порождать разные алгоритмы классификации. В данной работе эти понятия определяются для качественных признаков (измеренных в шкале наименований).

2. Пусть A - генеральная совокупность объектов и x_1, \dots, x_n - признаки, определенные на A , $I_i = \{x_i(a) | a \in A\} = \{x_{i1}, \dots, x_{ik(i)}\}$, $i=1, \dots, n$ - множества значений признаков. Определим атомарные одноместные отношения $P_{ij}(a) = x_i(a) = x_{ij}$, $x_{ij} \in I_i$, $i = 1, \dots, n$; $j = 1, \dots, k$. Множество атомарных отношений обозначим через X . **Детерминированной закономерностью** назовем истинную на A формулу вида $\forall a \Phi(a)$, составленную из атомарных отношений и их отрицаний.

ОПРЕДЕЛЕНИЕ 1. Импликативной детерминированной закономерностью назовем истинную на A формулу вида

$$F = \forall a (P_{i1j1}^{\varepsilon 1}(a) \& \dots \& P_{i1j1}^{\varepsilon 1}(a) \Rightarrow P_{i0j0}^{\varepsilon 0}(a)), \quad (1)$$

где $\varepsilon = 1(0)$, если отношение берется без отрицания (с отрицанием), удовлетворяющую следующим условиям:

- а) среди атомарных отношений $P_{i1j1}^{\varepsilon 1}(a), \dots, P_{i1j1}^{\varepsilon 1}(a), P_{i0j0}^{\varepsilon 0}$ нет повторений и нет одновременно отношения и его отрицания;
- б) если из конъюнкции $P_{i1j1}^{\varepsilon 1}(a) \& \dots \& P_{i1j1}^{\varepsilon 1}(a)$ удалить одно из отношений, либо заменить отношение $P_{i0j0}^{\varepsilon 0}$ на Л(ложь), то полученная формула станет ложной на A .

В [9] доказано, что любая детерминированная закономерность логически эквивалентна совокупности импликативных детерминированных закономерностей. Расширим множество детерминированных закономерностей, добавив закономерности, истинные с большой вероятностью. Предположим, что задана некоторая процедура случайного выбора объектов из генеральной совокупности A и определена вероятность отношений из X .

ОПРЕДЕЛЕНИЕ 2. Формулу $(P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1j1}^{\varepsilon 1} \Rightarrow P_{i0j0}^{\varepsilon 0})$ вида (1) назовем **вероятностной закономерностью** (на A), если

$$1) \mu(P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1}) > 0;$$

$$2) \mu(P_{i0j0}^{\varepsilon 0} / P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1}) > \mu(P_{i0j0}^{\varepsilon 0} / P_{i1j1}^{\varepsilon 1} \& \dots \wedge \dots \wedge \dots \& P_{i1jl}^{\varepsilon 1}),$$

где $\dots \wedge \dots \wedge \dots$ означает отсутствие одного или нескольких отношений в конъюнкции;

3) При добавлении к конъюнкции $P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1}$ любого отношения из X (или его отрицания) нарушается одно из первых двух условий.

Обозначим условную вероятность $\mu(P_{i0j0}^{\varepsilon 0} / P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1})$ формулы F через $\mu(F)$.

ЛЕММА 1. Импликативные детерминированные закономерности являются вероятностными закономерностями.

ДОКАЗАТЕЛЬСТВО: Так как импликативная детерминированная закономерность перестает быть истинной на A при замене отношения $P_{i0j0}^{\varepsilon 0}$ на L , то конъюнкция $P_{i1j1}^{\varepsilon 1}(a) \& \dots \& P_{i1jl}^{\varepsilon 1}(a)$ не всегда ложна на A . Следовательно, $\mu(P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1}) > 0$.

Для импликативной детерминированной закономерности F $\mu(F) = 1$. Так как F перестает быть истинной на A при удалении каких-либо отношений из $\{P_{i1j1}^{\varepsilon 1}, \dots, P_{i1jl}^{\varepsilon 1}\}$ то $\mu(P_{i0j0}^{\varepsilon 0} / P_{i1j1}^{\varepsilon 1} \& \dots \wedge \dots \wedge \dots \& P_{i1jl}^{\varepsilon 1}) < 1$, откуда вытекает второе условие. При добавлении к конъюнкции $P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1}$ любого отношения $P \in X$ (или его отрицания) она становится либо тождественно ложной на A , и тогда нарушается первое условие, либо из истинности $P \& P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1}$ следует истинность $P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1}$, а значит и истинность $P_{i0j0}^{\varepsilon 0}$. Откуда следует, что условная вероятность $\mu(P_{i0j0}^{\varepsilon 0} / P \& P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1})$ также равна 1, и, значит, нарушается второе условие. Лемма доказана.

Из леммы 1 следует, что понятие вероятностной закономерности является расширением понятия детерминированной закономерности. Вероятностные закономерности можно с некоторым доверительным уровнем α обнаруживать по выборкам из A . В [9] приведен метод обнаружения вероятностных закономерностей, использующий для проверки условия 2 точный критерий независимости Фишера для таблиц сопряженности признаков. Применяя этот метод с некоторым доверительным уровнем α получим множество формул F_α вида (1). Формулы из F_α будем называть закономерностями. Для каждой закономерности $F \in F_\alpha$ и некоторого доверительного уровня β в методе [9] определяется нижняя доверительная граница $\mu^\beta(F)$ условной вероятности $\mu(F)$.

3. Определим наборы, являющиеся **идеальными описаниями классов**. Для этого введем критерий взаимной согласованности подтверждающихся и опровергающихся на этих наборах закономерностей.

Набором значений признаков x_{s1}, \dots, x_{sm} (признаки не повторяются) будем называть множество $\{Y_{s1}, \dots, Y_{sm}\}$, $Y_{st} \in I_{st}$, $Y_{st} \neq \emptyset$, $t = 1, \dots, m$. Каждый набор $\{Y_{s1}, \dots, Y_{sm}\}$ выделяет в произвольном подмножестве объектов $B \subset A$ подмножество $M_B(\{Y_{s1}, \dots, Y_{sm}\}) = \{a \in B \mid a = \langle x_{1j1}, \dots, x_{njn} \rangle, x_{stj(st)} \in Y_{st}, t = 1, \dots, m\}$. Будем говорить, что закономерность $(P_{i1j1}^{\varepsilon 1} \& \dots \& P_{i1jl}^{\varepsilon 1} \Rightarrow P_{i0j0}^{\varepsilon 0})$ применима к набору $\{Y_{s1}, \dots, Y_{sm}\}$, если $\{i_0, i_1, \dots, i_l\} \subset \{s_1, \dots, s_m\}$ и $x_{itjt} \in Y_{it}$ при $\varepsilon_t = 1$ и $(x_{itjt} \notin Y_{it})$ при $\varepsilon = 0$, $t = 1, \dots, l$ (заметим, что $t = 0$ отсутствует). Если закономерность применима к набору $\{Y_{s1}, \dots, Y_{sm}\}$ и ее заключение $P_{i0j0}^{\varepsilon 0}$ выполнимо на этом наборе: $x_{i0j0} \in Y_{i0}$ при $\varepsilon = 1$ и $x_{i0j0} \notin Y_{i0}$ при $\varepsilon = 0$, то будем говорить, что эта закономерность **подтверждается** на этом наборе. Если закономерность применима к набору, но ее заключение не выполняется на этом наборе: $x_{i0j0} \notin Y_{i0}$ при $\varepsilon = 1$ и $x_{i0j0} \in Y_{i0}$ при $\varepsilon = 0$, то будем говорить, что она **опровергается** на этом наборе.

Критерий взаимной согласованности закономерностей по предсказанию значений признаков объектов класса определим следующим образом:

$$\Gamma(\{Y_1, \dots, Y_m\}) = - \left(\sum_{\varphi \in \Pi} \ln(1 - \mu^{\beta}(\varphi)) - \sum_{\varphi \in O} \ln(1 - \mu^{\beta}(\varphi)) \right),$$

где Π - множество закономерностей из F_{α} , подтверждающихся на наборе $\{Y_1, \dots, Y_m\}$, O - множество закономерностей из F_{α} , опровергающихся на этом наборе.

ОПРЕДЕЛЕНИЕ 3. Идеальным представителем (эталонном) класса (образа, таксона) будем называть такой набор значений признаков $\{Y_1, \dots, Y_m\}$, для которого критерий Γ имеет локальный максимум: при изменении любого из множеств Y_1, \dots, Y_m на один элемент значение критерия строго уменьшается.

Эталоны классов образуют иерархию: эталон класса $\{Y_1, \dots, Y_m\}$ является более общим, чем эталон класса $\{Y'_1, \dots, Y'_m, Y'_{m+1}, \dots, Y'_M\}$, $Y'_i \subset Y_i$, $i = 1, \dots, m$.

4. Классификация. Возьмем выборку B из A и получим на ней множество закономерностей F^B_{α} . Определим все эталоны классов. Каждому эталону $\{Y_1, \dots, Y_m\}$ в выборке B соответствует подмножество $M_B(\{Y_1, \dots, Y_m\})$, которое назовем классом. Множество всех классов образует классификацию. Критерий Фишера достаточно чувствителен, чтобы позволить обнаруживать закономерности, выделяющие класс, состоящий из одного объекта (если этот объект обладает своеобразным сочетанием значений признаков). В выборке B могут быть объекты не входящие ни в один класс.

Объединение всех множеств Y_1, \dots, Y_m всех эталонов классов $\{Y_1, \dots, Y_m\}$ дает информативную систему значений признаков.

5. Распознавание. Пусть B - выборка из A и b - новый случайно выбранный из A объект. Обнаружим на множестве $B \cup b$ закономерности $F^{B \cup b}_{\alpha}$. Определим все эталоны классов и проведем классификацию объектов $B \cup b$. Возможны три случая:

- 1) объект b входит в некоторые классы, содержащие также объекты выборки B .
- 2) объект b составляет одноэлементный класс;
- 3) объект b не входит ни в один класс выборки B .

В собственном смысле распознавание, как соотнесение объекта b с другими объектами выборки, происходит только в первом случае. Во втором случае обнаруживается уникальный класс (аналогов и названия для такого метода не известны). В третьем случае определяется, что с точки зрения имеющихся признаков и закономерностей данный объект «случаен» (такие методы так же неизвестны).

6. Предсказание. Пусть для объекта b нужно предсказать одно или несколько неизвестных значений признаков. Проведем распознавание объекта b , как описано в предыдущем пункте. Это возможно, так как метод [9] обнаруживает закономерности при наличии пробелов. Объект $b = \langle x_{ij1}, \dots, \dots, x_{ijn} \rangle$ с пропущенными значениями признаков будем относить к классу $M_{B \cup b}(\{Y_{s1}, \dots, Y_{sm}\})$, если выполнены условия $x_{stj(st)} \in Y_{st}$, $t = 1, \dots, m$ для всех $x_{stj(st)}$, определенных в b . Если объект b не входит ни в один класс, то предсказание не делается. Если объект b составляет самостоятельный класс $b = M_{B \cup b}(\{Y^b_{s1}, \dots, Y^b_{sm}\})$, то для неопределенных признаков объекта b множества Y^b_{st} должны быть пусты. Поэтому предсказание в этом случае невозможно. Пусть объект b входит в классы $M^1_{B \cup b}(\{Y^1_{s1}, \dots, Y^1_{s1m}\}), \dots, M^k_{B \cup b}(\{Y^k_{sk1}, \dots, Y^k_{skm}\})$ и нам надо предсказать неизвестное значение признака x_s . Так как объекты каждого класса подчиняются своей группе закономерностей, то эти закономерности будут л о к а л ь н о - к о м п е т е н т н ы [10] для предсказания неизвестных значений признаков объектов данного класса. Если в некотором классе $M^i_{B \cup b}$ есть множество значений Y^i_s признака x_s то для объекта b предсказываются значения Y^i_s .

Если ни один из классов $M^1_{\text{вУб}}, \dots, M^k_{\text{вУб}}$ не содержит значений признака x_s , то предсказание не делается. Пусть i_1, \dots, i_k - номера классов, имеющих значения признака x_s . Тогда предсказанием по всем классам будет множество значений $Y_s = Y^{i_1}_s \cap \dots \cap Y^{i_k}_s$. Если множество Y_s не пусто, то оно является искомым предсказанием, если пусто, то - предсказание по разным классам противоречиво, и не осуществляется. Такой метод предсказания, учитывающий локальную компетентность классов, наличие требуемой для предсказания информации и отсутствия противоречий, является оригинальным и аналогов не имеет.

7. Замечания. Если, как и в [9], использовать закономерности, включающие произвольные многоместные отношения, то определение класса можно распространить и на данные, измеренные в других шкалах.

ЛИТЕРАТУРА

1. Загоруйко Н.Г. Таксономия в анизотроном пространстве.- В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып.76). Новосибирск, 1978, с. 26-33.
2. Классификация и кластер/Под ред. Дж. Вэн Райзин, перев. под ред. Ю.И.Журавлева. - М.:Мир, 1980. - 389 с.
3. Загоруйко Н.Г. Методы распознавания и их применение. -М.: Сов.радио, 1972 - 206 с.
4. Дуда Р., Харт П. Распознавание образов и анализ сцен -М.:Мир, 1976. - 510 с.
5. Ту Дж., Гонсалес Р. Принципы распознавания образов. -М.: Мир, 1978. - 410 с.
6. Забродин В.Ю. О критериях естественности классификаций.- НТИ, сер.2, 1981,
7. Шрейдер Ю.А. Шаров А.А. Системы и модели. М., Радио и Связь, 1982, с.152
8. ВИТЯЕВ Е.Е. Классификация, как выделение групп объектов, удовлетворяющих разным множествам согласованных закономерностей. В кн.: Анализ разнотипных данных (Вычислительные системы, вып.99), Новосибирск, 1983, с. 44-50.
9. Витяев Е.Е. Метод обнаружения закономерностей и метод предсказания. - В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып. 67), Новосибирск, 1976, с.54-68.
10. Загоруйко Н.Г., Елкина В.Н., Тимеркаев В.С. Алгоритмы заполнения пропусков в эмпирических таблицах (Алгоритм ZET). В кн.: Вычислительные системы, вып. 61, Новосибирск, 1975, с. 3-27.